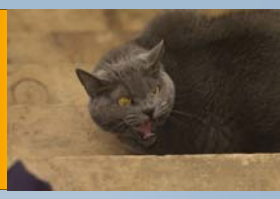
Distributions of gene tree branch lengths under coalescence

James H. Degnan, Department of Human Genetics, University of Michigan
Laura Salter Kubatko, Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University



BACKGROUND

Gene trees in species trees

- Nodes of gene trees represent coalescent events—ancestral gene copies.
- Genes from different species are constrained to coalesce in ancestral populations.
- Coalescent theory models coalescent events using exponential waiting times.
- Rannala and Yang (2003) describe the joint density of the coalescent times and gene tree topology. The form of the density depends on the *coalescent history* (Degnan and Salter, 2005), the list of populations in which coalescences occur.

Notation

- G is the random gene tree topology, σ is the species tree
- n taxa, s species, b is the index for the branch, c_b is the number of coalescences on branch b
- $\mathbf{h} = (h_1, h_2, \ldots, h_{n-1})$ is the coalescent history. Node i coalesces in population h_i
- τ_b species divergence time, N_b effective population size for population b
- $p(\tau_b)$ divergence time for node parental to b
- $u_b(\mathbf{h})$, $v_b(\mathbf{h})$ number of lineages "entering" and "exiting" population b
- $y_b(\mathbf{h})$ = time of final coalescence in population in population b or τ_b if $c_b(\mathbf{h}) = 0$
- $\mathbf{t} = (t_1, t_2, \ldots, t_{n-1})$ vector of coalescent times, t_{bi} = ith coalescence on branch b

Joint Density

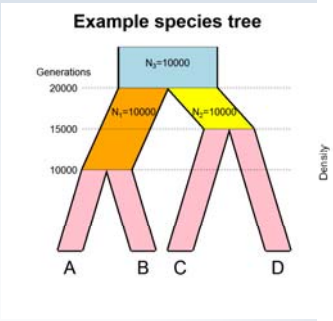
$$f_{\mathbf{T},G}(\mathbf{t},g) = \prod_{b=1}^{s-1} \exp\left\{-\binom{v_b(\mathbf{h})}{2}\frac{p(\tau_b)-y_b(\mathbf{h})}{2N_b}\right\} \times \prod_{i=1}^{c_b(\mathbf{h})} \frac{1}{N_b}\exp\left\{-\binom{u_b(\mathbf{h})-i+1}{2}\frac{t_{b_i}-t_{b_{i-1}}}{N_b}\right\}$$

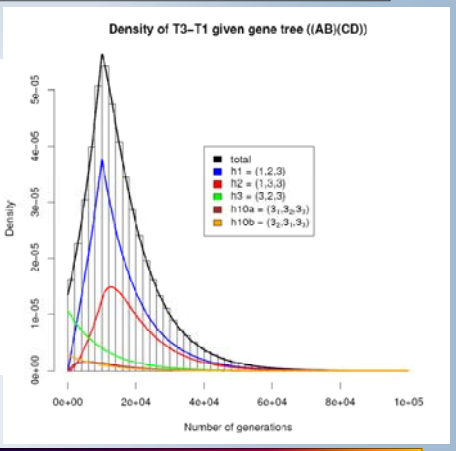
lineages not coalescing

coalescent events

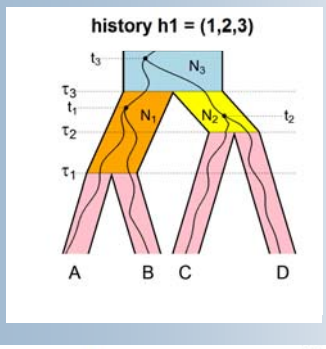
EXAMPLE BRANCH LENGTH DISTRIBUTION

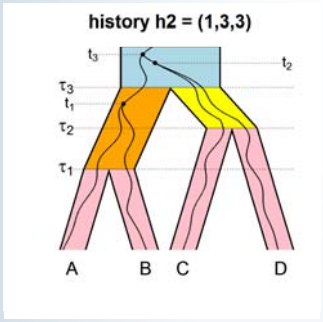
The example shows a particular species tree, from which 1,000,000 gene trees were simulated. From these were subset the 449,599 gene trees matching the species tree, and the length of the branch connecting the ancestor of A and B to the root was recorded. The histogram of simulated branch lengths is reported along with the theoretical density obtained by using the joint distribution of T_1 and T_3 and deriving the distribution of $U=T_3-T_1$.

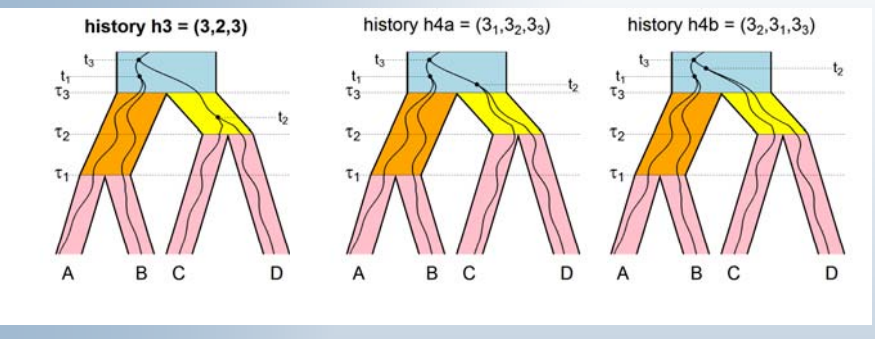




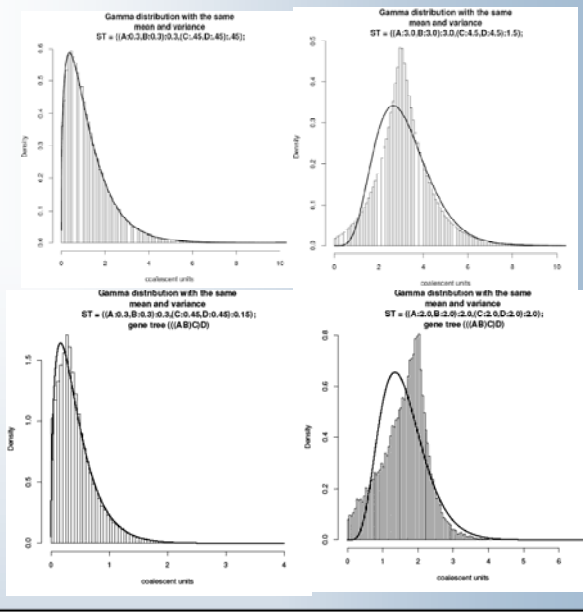
COALESCENT HISTORIES







CONCLUSIONS



- For moderately short branch lengths on the species tree, the distribution of a branch length for a gene trees often results from a mixture depending on which coalescent history occurs
- The distribution of a branch length given a species tree and gene tree is sometimes not approximately gamma distributed, although these are often used as priors for branch lengths in Bayesian phylogenetics.
- A gamma distribution can sometimes do a better job of approximating the branch length density when there is *MORE* lineage sorting due to short branches (see figure) and do a worse for moderate branch lengths. This pattern seems to hold when the gene tree does not match the species tree.

INTEGRATING JOINT DENSITIES USING COALESCENT HISTORIES

To find the distribution of the branch length $U=T_3-T_1$, we can use the joint distribution of T_1 and T_3, conditional on the gene tree topology. This joint density can be obtained by integrating the joint density of T_1, T_2, and T_3 with respect to T_2, conditional on the gene tree. This can be done by breaking up the region of integration over coalescent histories, and different coalescent histories suggest different limits of integration. For example, for h_1 and h_3, $\tau_2 < t_2 < \tau_3$, while for h_2 and h_4, $\tau_3 < t_2$.

To derive the density of $U=T_1-T_3$, let $V=T_1$. The joint density of U and V can be found using a bivariate transformation, and the distribution U can then be derived by integrating with respect to V. For notation, we let $f_{i,T1,T3}$ be the joint density of T_1 and T_3 when h_i is the coalescent history.

$$f_{U,V}(u,v) = f_{T_1,T_3}(t_1,t_3)/P(G=g\,|\,\sigma)$$
$$= f_{1,T_1,T_3}(v,u+v) + f_{2,T_1,T_3}(v,u+v) + f_{3,T_1,T_3}(v,u+v) + f_{4a,T_1,T_3}(v,u+v) + f_{4b,T_1,T_3}(v,u+v)$$
$$(1-(2/3)\exp(-(\tau_3-\tau_1)/N_1))(1-(2/3)\exp(-(\tau_3-\tau_2)/N_2))$$

BIBLIOGRAPHY

Degnan, JH and LA Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59: 24-37.

Rannala, B and Z Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.